DUKE
COMPUTER SCIENCE

CompSci 590.01
Spring 2023

Causal Inference in Data Analysis with Applications to Fairness and Explanations

Lecture 7:
Explanations
&
Causal Explanations

Sudeepa Roy

Reading

- Galhotra et al. et al., SIGMOD (2021): Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals (https://dl.acm.org/doi/10.1145/3448016.3458455)

Acknowledgement (big thanks!):

Slides are from a tutorial by Prof. Babak Salimi (UCSD), Prof. Romila Pradhan (Purdue), Dr. Sainyam Galhotra (U. Chicago), and graduate student Aditya Lahiri with small modifications from a tutorial they gave at ICDE'22 and SIGMOD'22, and from a joint short course at the Reasoning Web'22 summer school with Prof. Salimi

Announcements: Feb 9

- Please see presentation guideline on Ed
- Initial project feedback coming soon
- Send me an email by today with your interests if you are currently only yourself in a project team and looking for 1-2 teammates

Causal Inference and Explanations

Explanations – another important topic in Responsible Data Science

PURDUE UNIVERSITY UC San Diego THE UNIVERSITY OF CHICAGO

Explainable AI:
Foundations, Applications, and Opportunities for Data Management Research

Romila Pradhan
Assistant Professor
Purdue

Aditya Lahiri
Graduate Student
UCSD

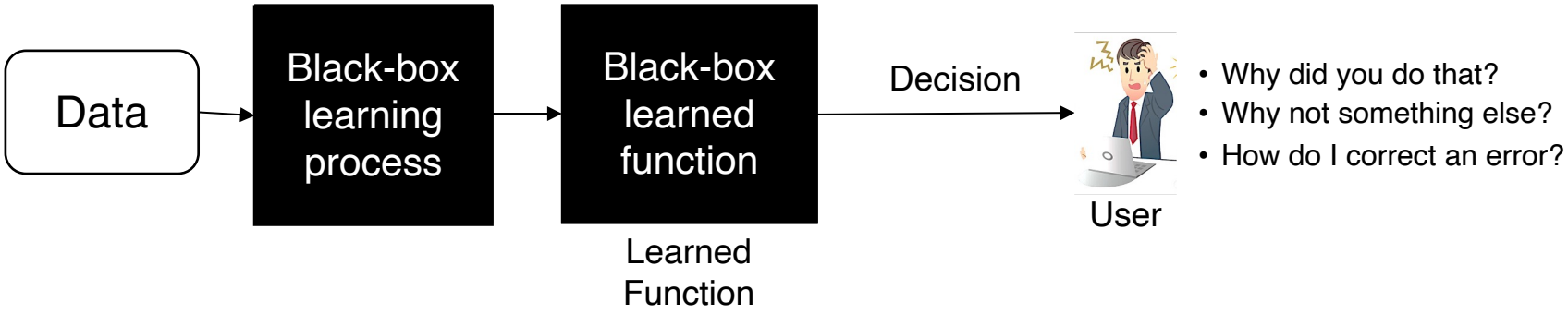
Sainyam Galhotra
Postdoc
UChicago

Babak Salimi
Assistant Professor
UCSD

https://explainable-ai-tutorial.github.io/
[SIGMOD'22, ICDE'22]

We interact with algorithmic decision-making on daily basis

Sophisticated ML models shown to be highly accurate for many applications



- Complex models → difficult to trace decisions back to questions about why and how they were made
- ML systems are often opaque– complexity, proprietary

Image source: DARPA XAI Program 2019

What is explainable AI?

- Human-understandable explanations of outcomes of algorithmic decision-making systems

- A powerful tool for answering How? and Why? questions about algorithmic systems – also for legal & ethical concerns

Not a new topic!

What is explainable AI?

HPP-79-11 (Working Paper) March 1979

Transfer of Expertise: A Theme for AI Research

A key idea in our current approach to building expert systems is that these programs should not only be able to apply the corpus of expert knowledge to specific problems, but they should also be able to interact with the users and experts just as humans do when they learn, explain, and teach what they know. We will show, as we review the major

What is explainable AI?

Explanation in Second Generation Expert Systems

William R. Swartout[1] and Johanna D. Moore[2]

[1] USC/Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292

[2] University of Pittsburgh
Computer Science Department and
Learning Research and Development Center
Pittsburgh, PA 15260

Abstract. What is needed for good explanation? This paper begins by considering some desiderata for expert system explanation. These desiderata concern not only the form and content of the explanations, but also the impact of explanation generation on the expert system itself—how it is built and how it performs. In this paper, we use these desiderata as a yardstick for measuring progress in the field. The paper describes

What is explainable AI?

A Theory of Diagnosis from First Principles

Raymond Reiter
Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 1A4; The Canadian Institute for Advanced Research

Recommended by Johan de Kleer and Daniel G. Bobrow

ABSTRACT

Suppose one is given a description of a system, together with an observation of the system's behaviour which conflicts with the way the system is meant to behave. The diagnostic problem is to determine those components of the system which, when assumed to be functioning abnormally, will explain the discrepancy between the observed and correct system behaviour.

We propose a general theory for this problem. The theory requires only that the system be described in a suitable logic. Moreover, there are many such suitable logics, e.g. first-order, temporal, dynamic, etc. As a result, the theory accommodates diagnostic reasoning in a wide variety of practical settings, including digital and analogue circuits, medicine, and database updates. The theory leads to an algorithm for computing all diagnoses, and to various results concerning principles of measurement for discriminating among competing diagnoses. Finally, the theory reveals close connections between diagnostic reasoning and nonmonotonic reasoning.

how it is built and how it performs. In this paper, we use these desiderata as a yardstick for measuring progress in the field. The paper describes

Applications of XAI

Prevent bias and discrimination

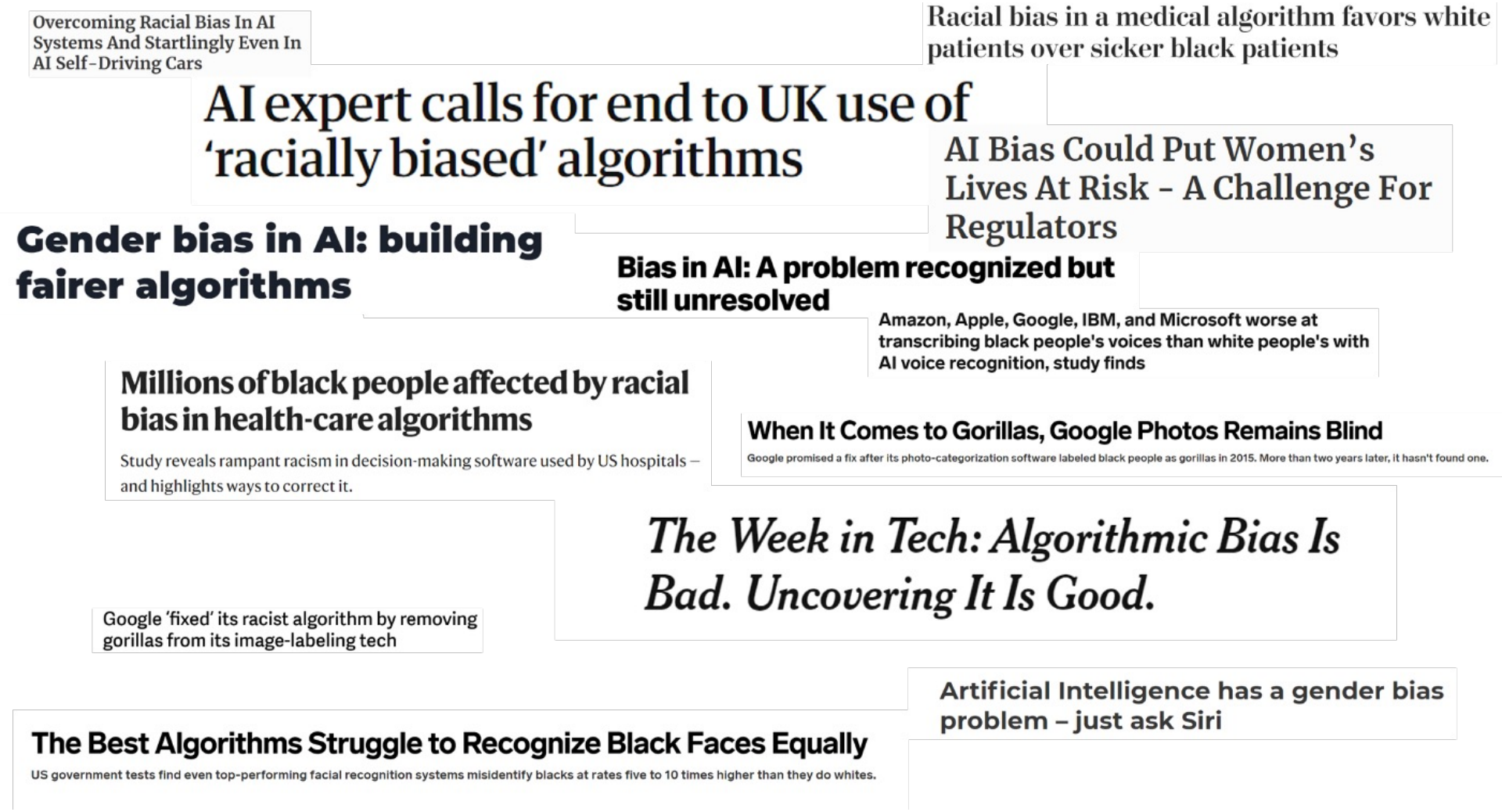
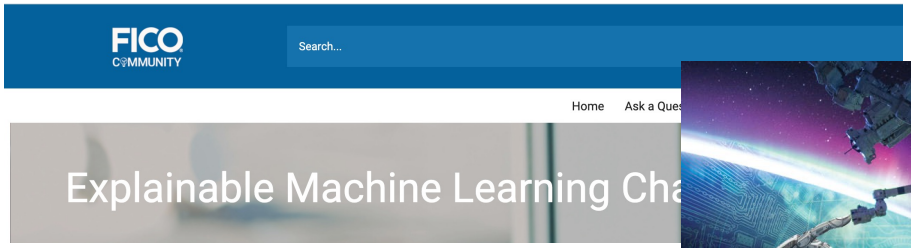


Image source: https://towardsdatascience.com/algorithm-bias-in-artificial-intelligence-needs-to-be-discussed-and-

Applications of XAI

Accuracy, trust, recourse and compliance with the law



"

Companies should commit to ensuring systems that could fall under GDPR, including AI, will be compliant. The threat of **sizeable fines of €20 million or 4% of global turnover** provides a sharp incentive.

Article 22 of GDPR empowers individuals with the **right to demand an explanation of how an AI system made a decision that affects them.**

"

European Commission

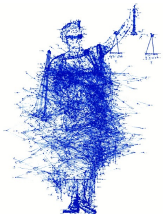
"The data subject shall have the **right not to be subject to a decision based solely on automated processing…**"

GDPR.EU

"… **monitor city use of algorithmic decision-making** and provide recommendations…"

NYC Automated Decision Systems Task Force

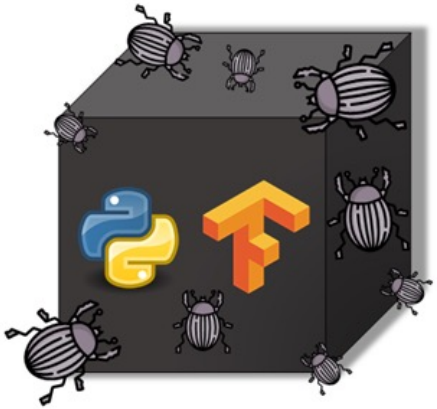
required companies to **study algorithms they use, identify bias in these systems and fix any discrimination** or bias they find

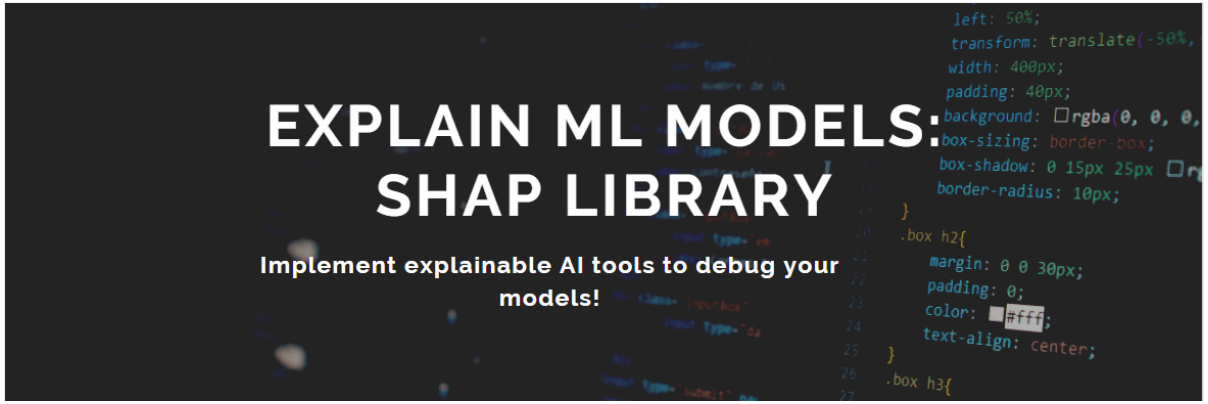


Algorithmic Accountability Act 2019

Applications of XAI

Debugging ML model

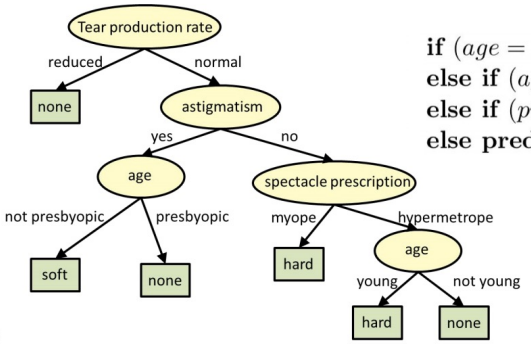




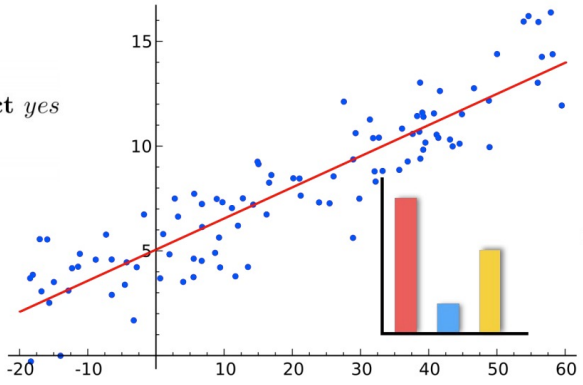
How are such decisions explained?

Covered in student presentations

1. Feature-based Explanations
2. Data-based Explanations

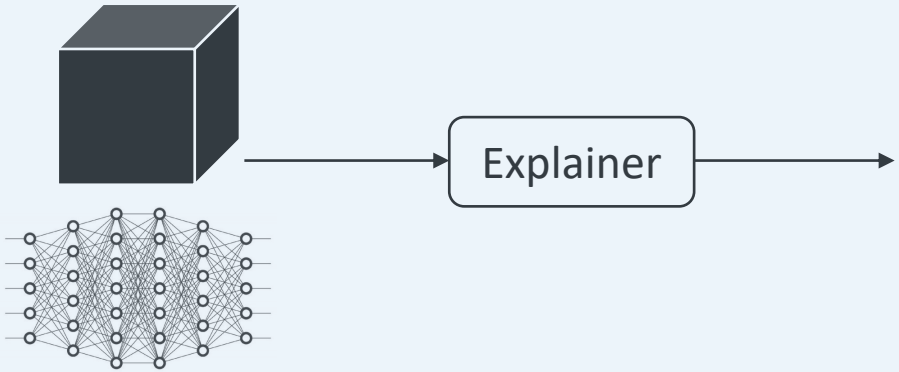


if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

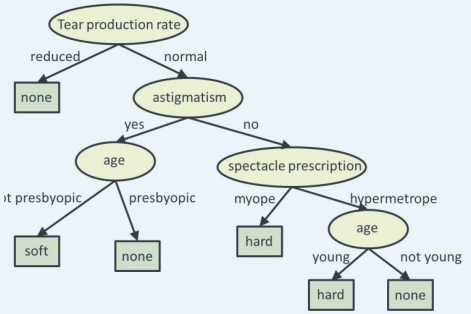


Intrinsic

Post hoc



if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no



Today

Feature-based explanations

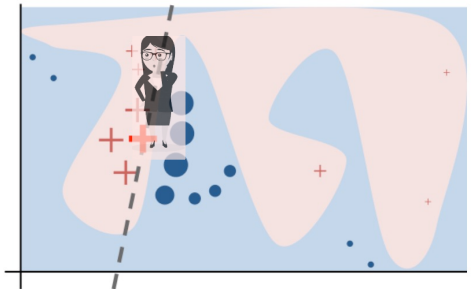
Feature Importance:
Identify important attributes and present their relative importance

Name	Age	Saving	Month	…	Credit History
Maeve	<25	<100 DM	24	…	Paid duly

ML Model

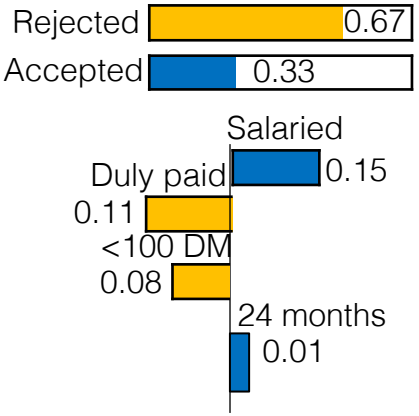
48%

REJECTED

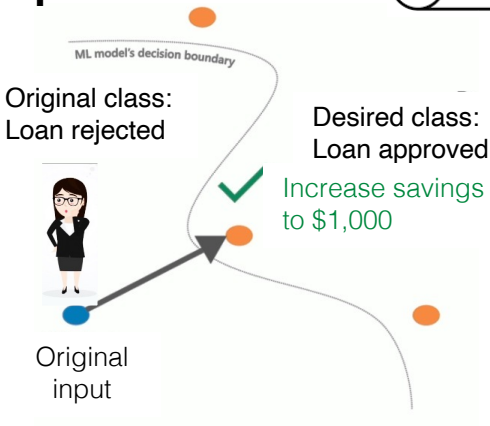


LIME. Identify important attributes to explain outcome for a single instance

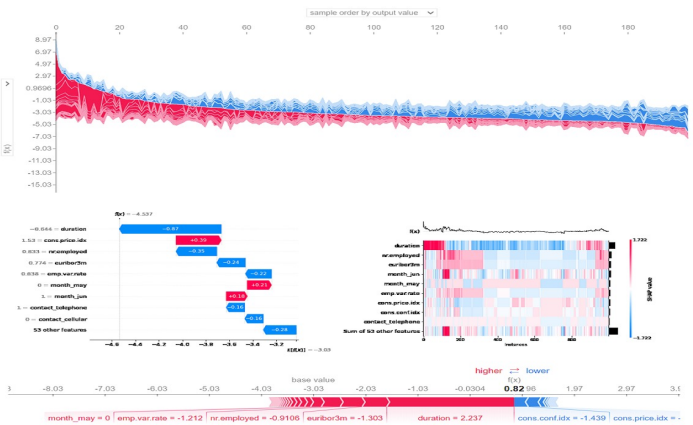
Source: Ribeiro et al
SIGKDD 2016



Counterfactual explanations:



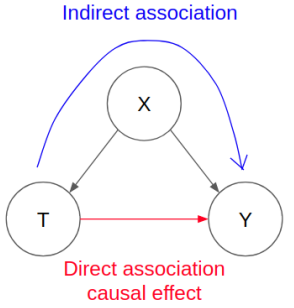
Source: Microsoft Research blog

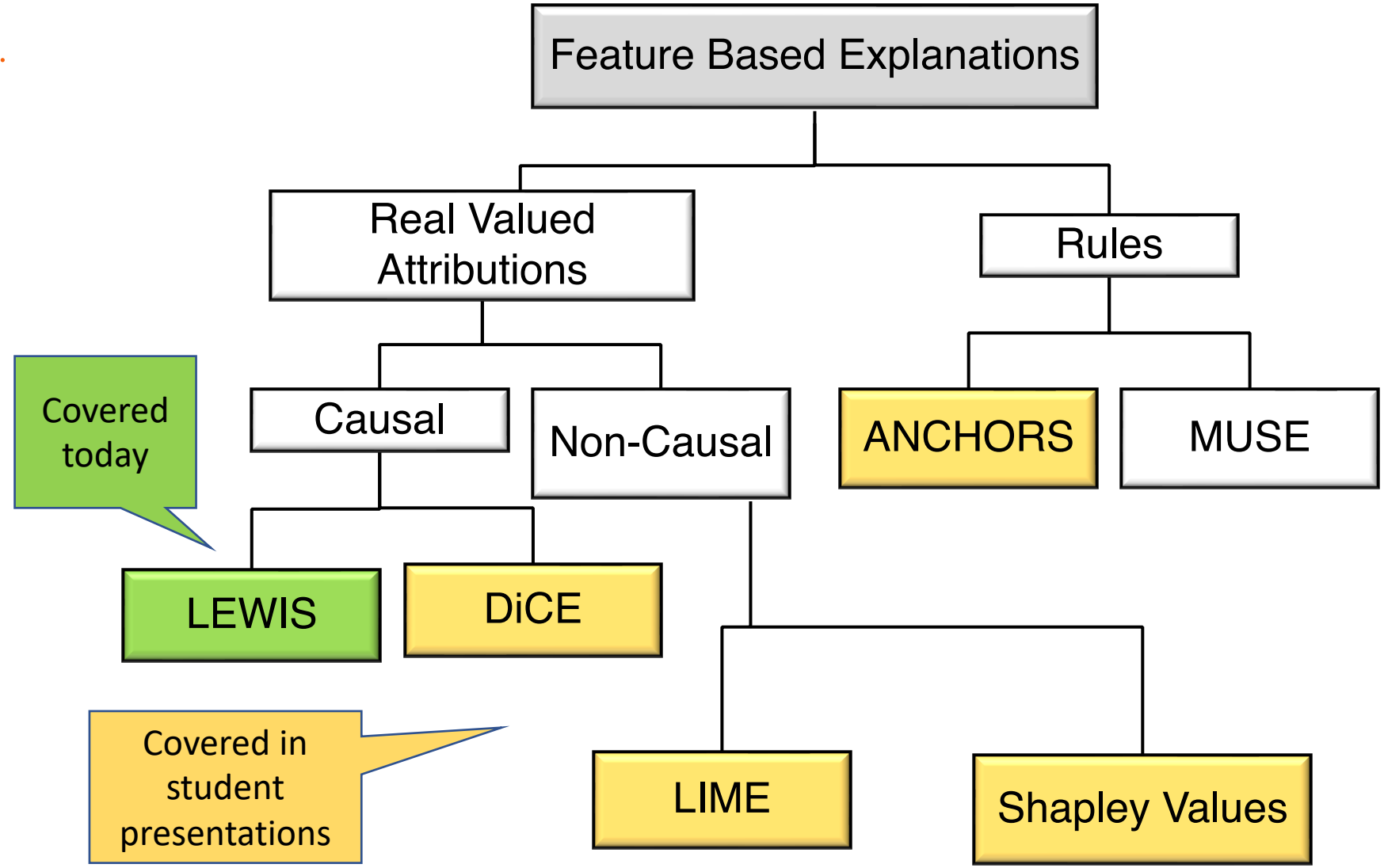


SHAP. Marginal contribution of each feature toward the prediction, averaged over all possible coalitions

Covered in student presentations

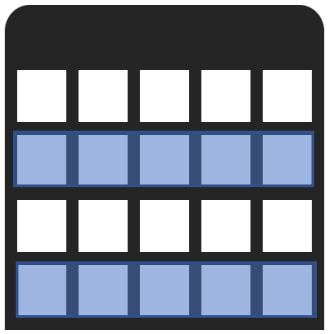
Causal explanations:



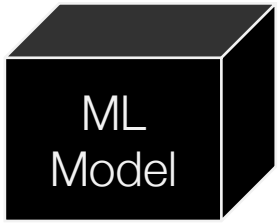

Feature Based Explanations
Real Valued Attributions
Rules
Causal
Non-Causal
ANCHORS
MUSE
Covered today
LEWIS
DiCE
Covered in student presentations
LIME
Shapley Values

Data-based explanations

- Feature-based explanations fall short in generating diagnostic explanations
- Data-based explanations: parts of the training data are responsible for an unexpected and discriminatory behavior of an ML model?



Training Data



Output

82.5%

Why this output ?

More later if we have time, otherwise check out the slides yourself ☺

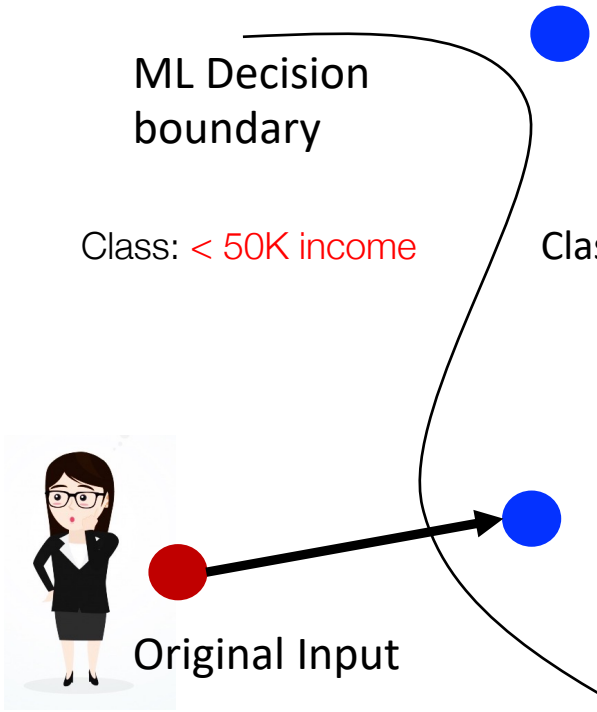
Causal Explanations

Counterfactual explanations

Name	Age	Saving	Month	…	Credit History
Maeve	<25	<100 DM	24	…	Paid duly

Counterfactual Explanations:

how the world would have (had) to be different for a desirable outcome to occur”



These are contrastive explanations

closely related: algorithmic recourse – the actions required for reversing unfavorable decisions across a range of counterfactual scenarios

Counterfactual Explanations

- Goal: Minimum change in input attributes to flip a model's prediction

- Popular Techniques
  - Feature-based explanations: Feature score captures its likelihood to change outcome
  - Instance-level explanations: Identify counterfactual scenarios to change outcome
    - Inverse Classification
    - Nearest-counterfactual explanation

Feature–based explanation

- Counterfactual score of a feature: Expected difference in outcome when the feature takes a value of any other individual value

Name	Age	Saving	Month	…	Credit Amount
Maeve	<25	<100 DM	24	…	Existing paid duly

Black-box ML model → REJECTED

Name	Age	Saving	Month	…	Credit Amount
Tom	<25	500 DM	24	…	Existing paid duly

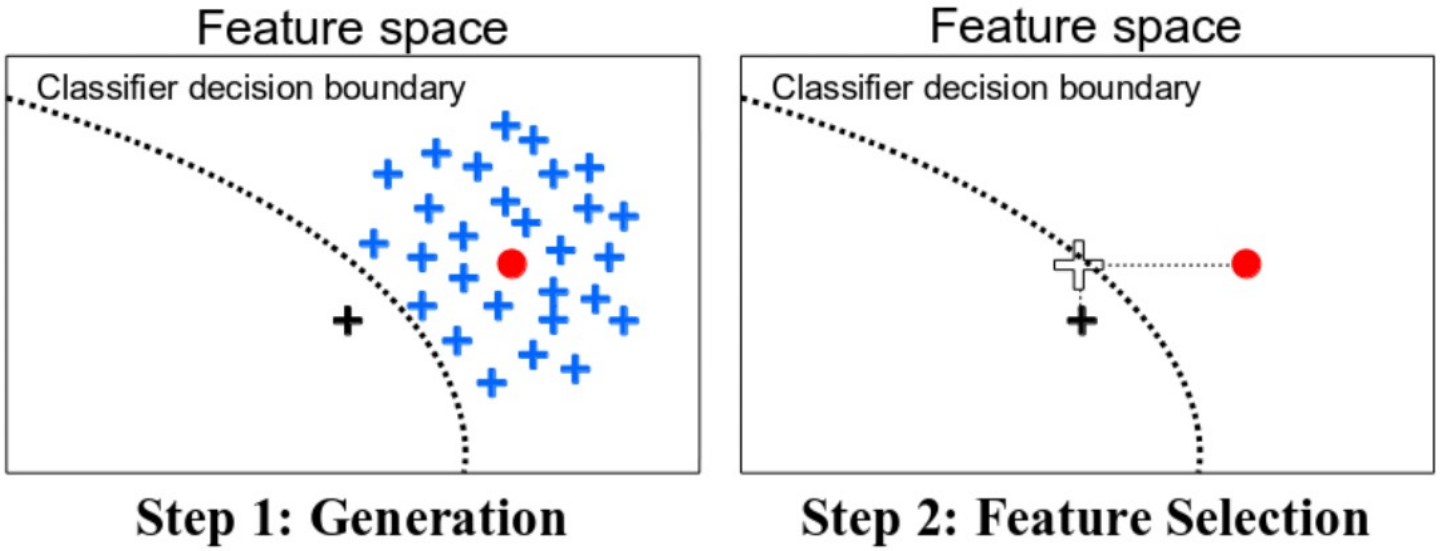
Black-box ML model → APPROVED

Name	Age	Saving	Month	…	Credit Amount
Tim	<25	200 DM	24	…	Existing paid duly

Black-box ML model → REJECTED

Inverse Classification

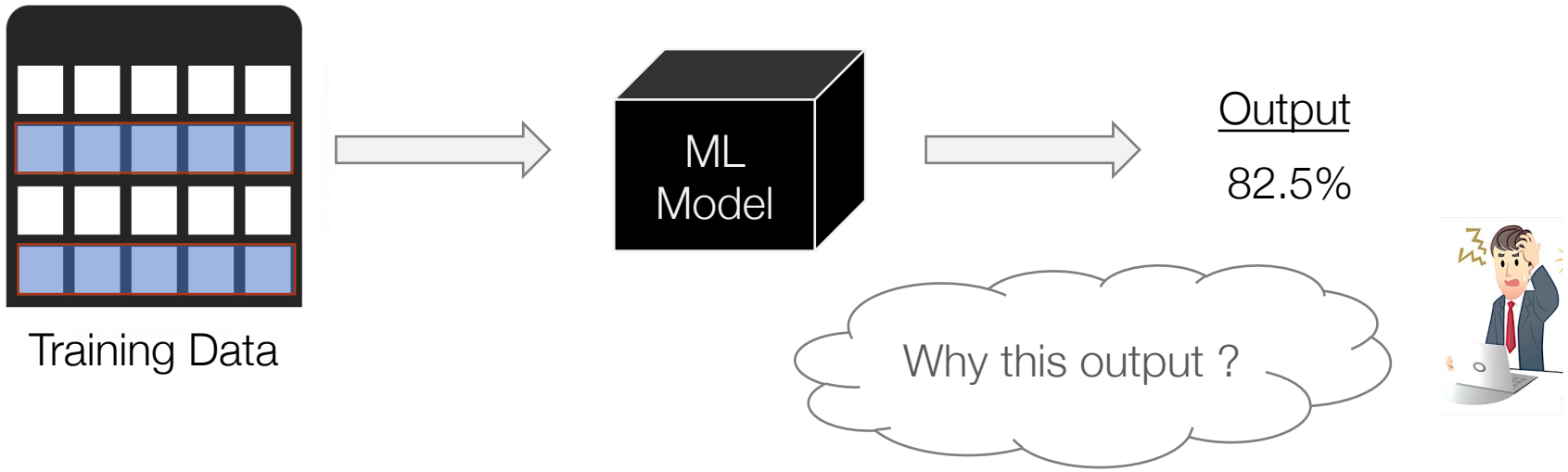
- Identify a close neighbor of the point which has the opposite outcome
- Growing sphere algorithm



Inverse Classification for Comparison-based Interpretability in Machine Learning
Laugel et al., https://arxiv.org/pdf/1712.08443.pdf - figure from the paper

Limitations

- Focus on correlation between input/output
- Fail to account for the causal interaction between attributes
- Perturbations not translatable to real-world interventions
- Fall short in generating diagnostic explanations



What kind of explanations do humans seek?

CONTRASTIVE

"Explanatory relevant information is information that is potentially **relevant to manipulation and control**"
—James Woodward, Philosopher

The key insight is to recognize that one does not explain events per se, but that one explains **why the puzzling event occurred in the target cases but not in some counterfactual contrast case**.
—Denis Hilton, Psychologist

"An explanation is an **assignment of causal responsibility**"
—Josephson and Josephson, Computer Scientists

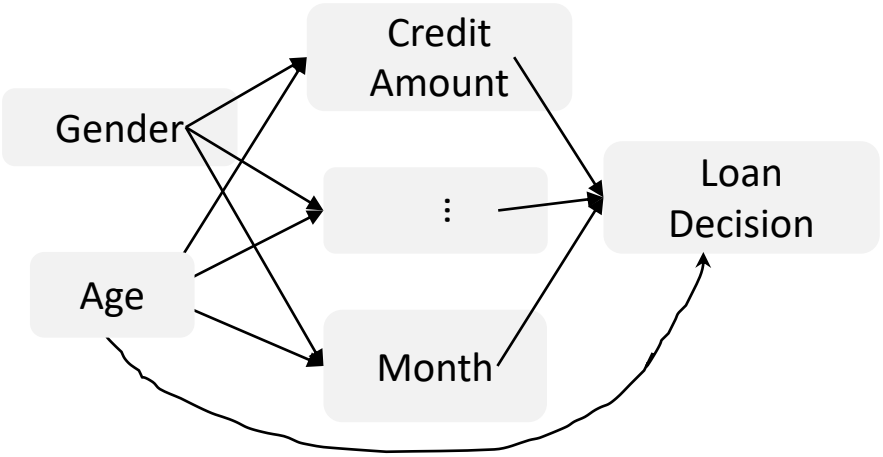
"To explain an event is to provide some **information about its causal history**."
"… think of a cause as something that makes a difference… **Had it been absent, its effects** – some of them, at least, and usually all – **would have been absent as well**" —David Lewis, Philosopher

CAUSAL

"…the only way to **ensure that the recommended change is even possible**…"
"…and to **account for dependencies between features** is to model the outcome of interest using features that directly figure into the causal mechanism"
—Barocas et al. FAT 2020

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1-38.

Recap: Causal Graph



Fully specified model (functions and values of exogenous variables known)

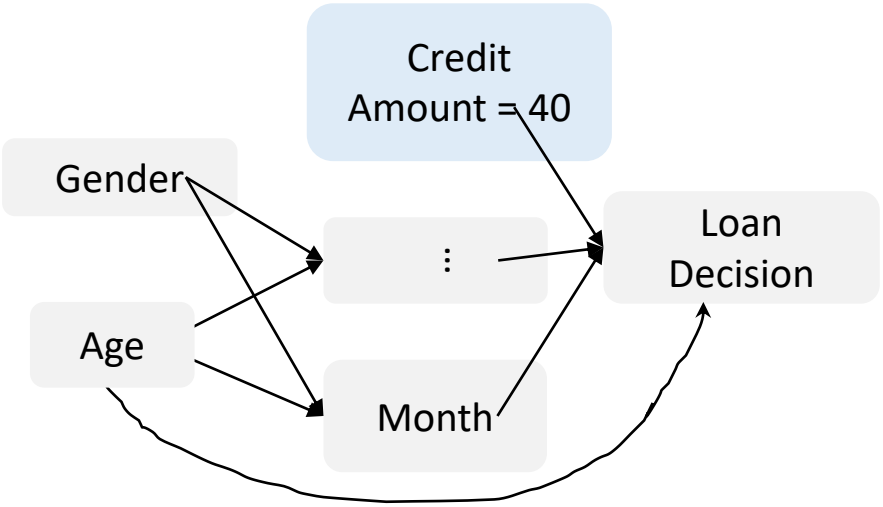
$Age = U_{Age}$

$Gender = U_{Gender}$

$Month = 0.4 \cdot Age + 0.8 \cdot Gender + U_{Month}$

$Decision = Month + 0.5 \cdot Age + U_{Decision}$

Recap: Counterfactuals/Interventions



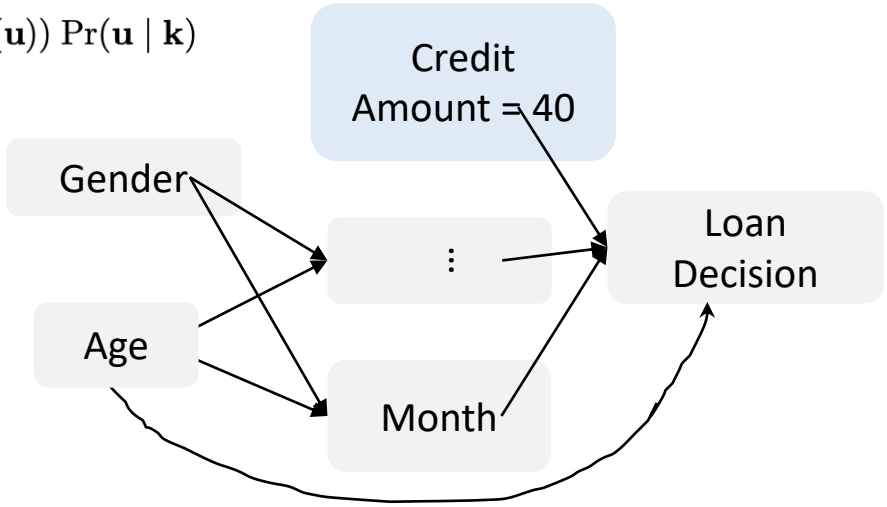
Equivalent to removing incoming edges of credit amount and conditioning on it

$$\Pr(y_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{k}) = \sum_{\mathbf{u}} \Pr(y_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}))\ \Pr(\mathbf{u} \mid \mathbf{k})$$

Pr(u) induces a probability distribution over exogenous variables

Recap: Counterfactuals/Interventions

$$\Pr(y_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{k}) = \sum_{\mathbf{u}} \Pr(y_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}))\ \Pr(\mathbf{u} \mid \mathbf{k})$$



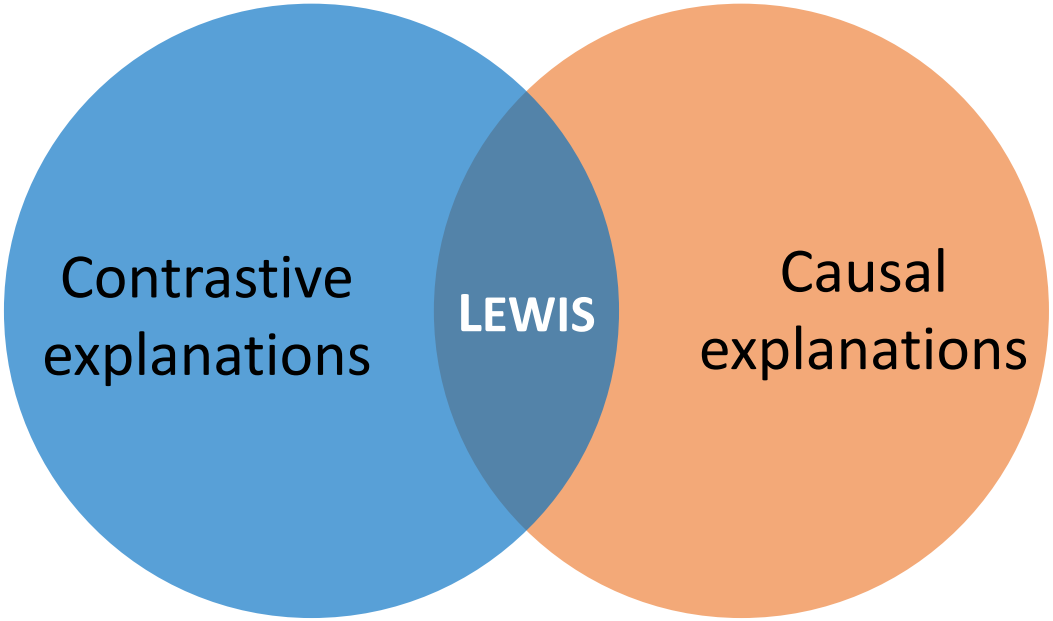
Pearl's three step procedure

1. Abduction: Calculate Pr(u|k)
2. Action: modify the causal graph to reflect intervention
3. Prediction: Calculate the probability of Y=y

Causal Explanations

- Advantages
  - Translatable to real-world interventions
  - Easily understandable
  - Robust against spurious correlations in the data
- Challenges
  - Estimation requires additional information
    - Domain knowledge
    - Causal graph

Lewis: Unifying Contrastive and feature attribution methods



Based on **probabilistic contrastive counterfactuals**

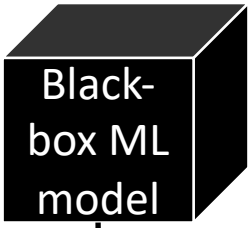
For individual(s) with attribute(s) <actual-value> for whom an algorithm made the decision <actual-outcome>, the decision would have been <foil-outcome> with probability <score> had the attribute been <counterfactual-value>.

To what extent is an attribute sufficient?

Rosa

Name	Age	Saving	Month	…	Credit Amount
Maeve	<25	<100 DM	**36**	…	Existing paid duly

Maeve's data



Rosa's credibility **87%**

What is the probability that loan would be approved if month **had been** 36 months instead of 24 months?

To what extent is an attribute necessary?

Irrfan,
a bank customer

Name	Age	Saving	Month	…	Credit History
Irrfan	>25	>500 DM	24	…	Existing paid duly

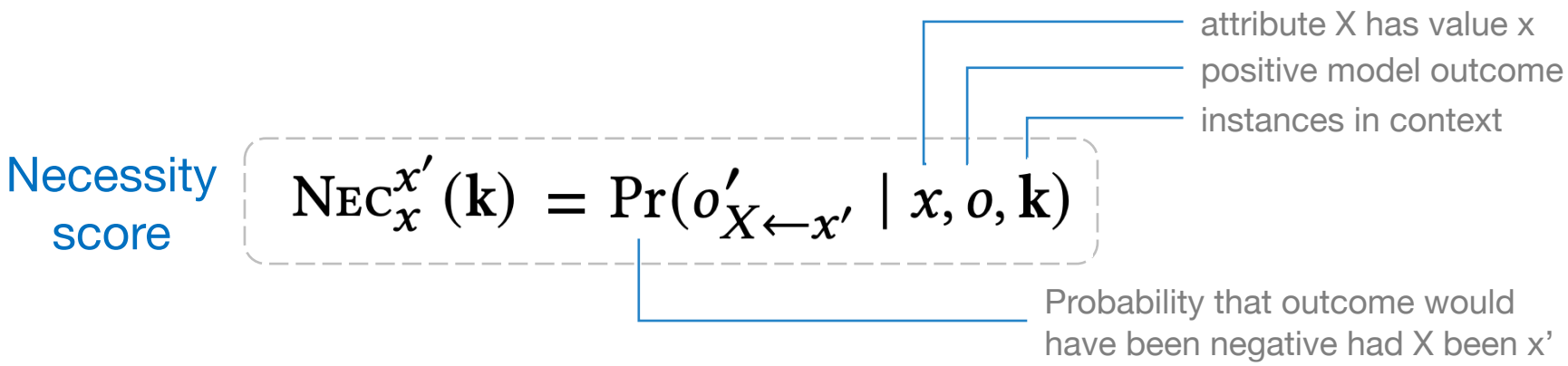
Irrfan's data

Black-box ML model

Irrfan's credibility 48%

What is the probability that loan would be rejected if month had been 24 months instead of 36 months?

Our explanation scores



Probability that
Loan=Rejected if Month were 24
for instances in context for whom
Loan=Approved when Month=36

Our explanation scores

$$\text{NEC}_x^{x'}(\mathbf{k}) = \Pr(o'_{X \leftarrow x'} \mid x, o, \mathbf{k})$$

$$\text{SUF}_x^{x'}(\mathbf{k}) = \Pr(o_{X \leftarrow x} \mid x', o', \mathbf{k})$$

attribute X has value x'

negative model outcome

instances in context

Probability that outcome would have been positive had X been x

Sufficiency score

Probability that
Loan=Approved if Month were 36
for instances in context for whom
Loan=Rejected when Month=24

Lewis

		For context k
$\text{NEC}_x^{x'}(\mathbf{k})$	$\Pr(o'_{X \leftarrow x'} \mid x, o, \mathbf{k})$	Probability that Loan=Rejected were Month=24 for whom Loan=Approved when Month=36
$\text{SUF}_x^{x'}(\mathbf{k})$	$\Pr(o_{X \leftarrow x} \mid x', o', \mathbf{k})$	Probability that Loan=Approved were Month=36 for whom Loan=Rejected when Month=24
$\text{NESUF}_x^{x'}(\mathbf{k})$	$\Pr(o_{X \leftarrow x}, o'_{X \leftarrow x'} \mid \mathbf{k})$	Probability that Loan=Approved were Month=36 and Loan=Rejected were Month=24

Our explanation scores

$$\text{NEC}_x^{x'}(\mathbf{k}) = \Pr(o'_{X \leftarrow x'} \mid x, o, \mathbf{k})$$

$$\text{SUF}_x^{x'}(\mathbf{k}) = \Pr(o_{X \leftarrow x} \mid x', o', \mathbf{k})$$

$$\text{NESUF}_x^{x'}(\mathbf{k}) = \Pr(o_{X \leftarrow x}, o'_{X \leftarrow x'} \mid \mathbf{k})$$

$$\begin{aligned}\text{NESUF}_x^{x'}(\mathbf{k}) \leq \quad & \Pr(o, \mathbf{x} \mid \mathbf{k})\, \textit{NEC}_{\mathbf{x}}(\mathbf{k}) \\ & + \Pr(o', \mathbf{x}' \mid \mathbf{k})\, \textit{SUF}_{\mathbf{x}}(\mathbf{k}) \\ & + 1 - \Pr(\mathbf{x} \mid \mathbf{k}) - \Pr(\mathbf{x}' \mid \mathbf{k})\end{aligned}$$

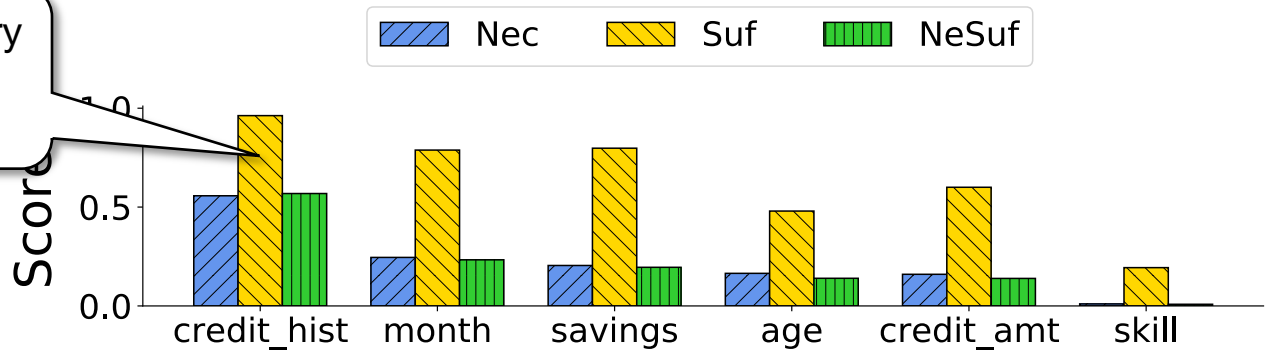
NeSuf can be seen to be a linear combination of Nec and Suf scores,
And for binary X, this inequality becomes an equality and NeSuf can be seen as the weighted average of Nec and Suf scores.

From scores to explanations

$$\textsc{Suf}_x^{x'}(\mathbf{k})$$

Global explanations

An improvement in credit history is most likely to change a negative decision into positive.

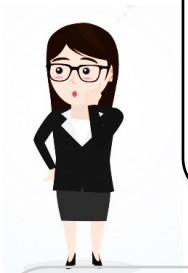


$$\text{SUF}_{x}^{x'}(\mathbf{k} = \phi)$$

- Scores express the global influence of attributes on algorithm's decision
- Maximum score over all pairs of an attribute's values

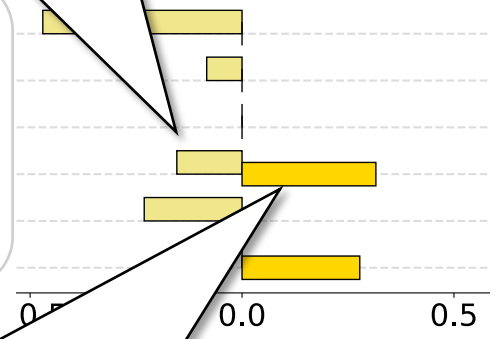
Local explanations

$$\textsc{Suf}_x^{x'}(\mathbf{k} = \text{Maeve's data})$$



Changing from current credit history value to a better value would have resulted in positive outcome with 15% probability.

Age	< 25 years
Credit amount	1,275 DM
Month	10 months
Credit history	Existing paid duly
Savings	< 100 DM
Skill level	Skilled



- Contribution of attri oward

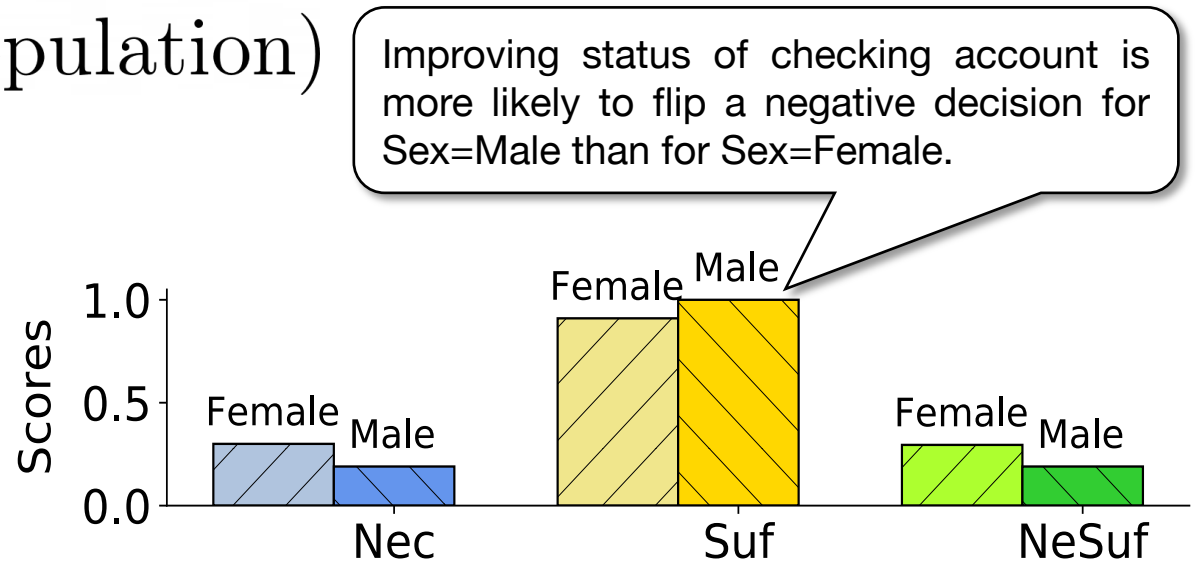
Staying at current credit history value instead of a worse value has positive contribution toward outcome.

- Positive and negative contribution of an attribute value e.g., credit history

Contextual explanations

$$\textsc{Suf}_x^{x'}(\mathbf{k} = \text{Sub-population})$$

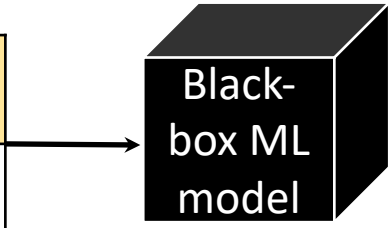
Effect of changing
Status of checking account



Algorithmic Recourse

- Input: Individual with negative outcome
- Goal: Identify intervention with smallest cost that can flip the outcome in the future

Name	Age	Saving	Month	…	Credit Amount
Maeve	<25	<100 DM	24	…	Existing paid duly



What should I do to flip the outcome?

Non-causal methods

- Model recourse as an optimization problem
- Linear Program:
  - Objective: Minimize the change in attributes
  - Constraints: Flip model outcome

- Assumption: All attributes are independent and do not impact each other

Ustun et al., Actionable Recourse in Linear Classification, ACM FAT ‘19

Causal Techniques

- Formulate as an optimization problem
- Goal: Minimize the cost of actions
- Constraints:
  - Classifier outcome is flipped
  - Modified input is causally consistent with the original input
- Assumption: Requires fully specified causal model

Karimi et al., Algorithmic Recourse: from Counterfactual Explanations to Interventions ACM FAacT'20

Lewis: How can I improve my chances of getting a loan?

User provides a set of actionable attributes A

Name	Age	Saving	Month	…	Credit History
Maeve	<25	<100 DM	24	…	Existing paid duly

Counterfactual recourse as interventions over actionable attributes

$$\underset{\mathbf{a}\in \mathrm{Dom}(\mathbf{A})}{\operatorname{argmin}} \quad \mathrm{Cost}\,(\mathbf{a}, \hat{\mathbf{a}})$$

$$\text{s.t.} \quad \textsc{Suf}_{\hat{\mathbf{a}}}(\text{Maeve's data}) \geq 85\%$$

Recommended Recourse:

Actionable Attributes	Current Value	Required Value
Credit amount	1,275 DM	3,000 – 5,000 DM
Savings	< 100 DM	500 – 1,000 DM

This recourse will lead to a positive decision

Can operate at sub-population level

Requires causal structure but not dependence equations

Computational Challenges

$\text{NEC}_x^{x'}(\mathbf{k})$	$\Pr(o'_{X \leftarrow x'} \mid x, o, \mathbf{k})$
$\text{SUF}_x^{x'}(\mathbf{k})$	$\Pr(o_{X \leftarrow x} \mid x', o', \mathbf{k})$
$\text{NESUF}_x^{x'}(\mathbf{k})$	$\Pr(o_{X \leftarrow x}, o'_{X \leftarrow x'} \mid \mathbf{k})$

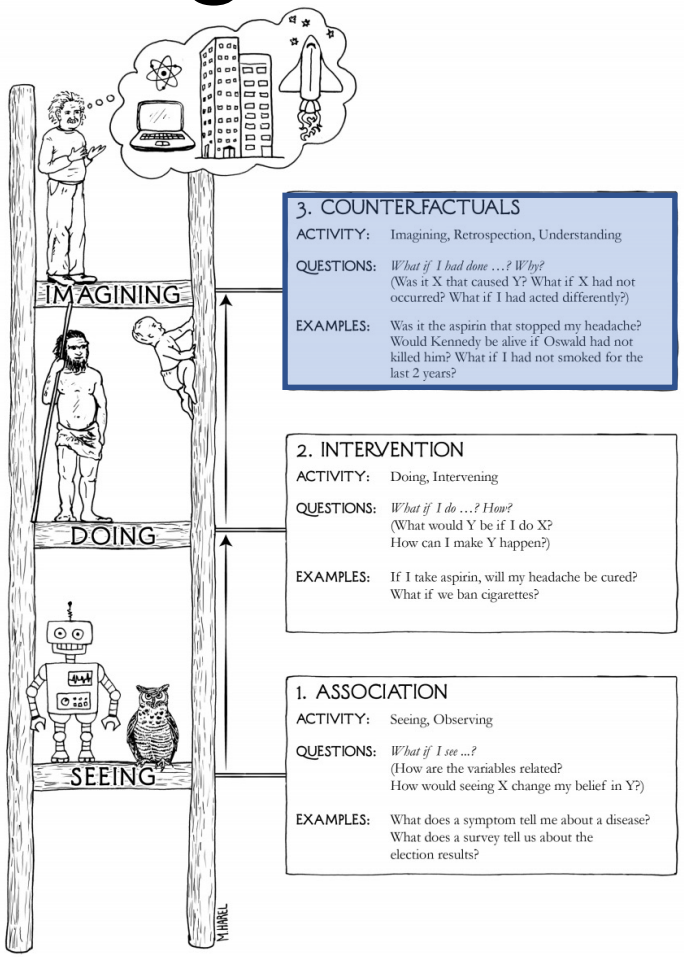
Requires the full structural causal model

Age = U_{Age}

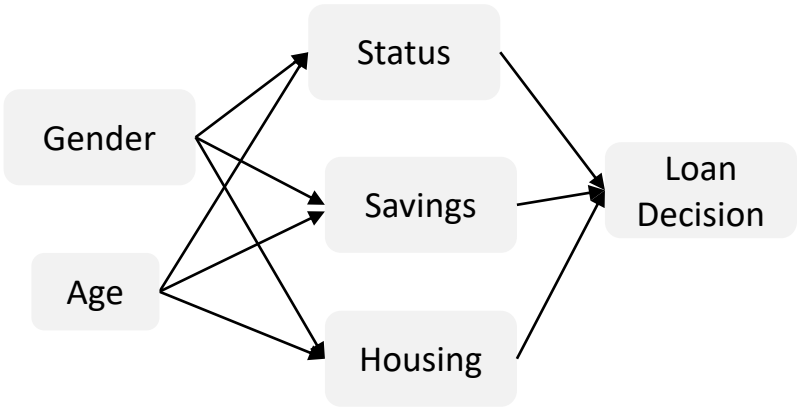
Gender = U_{Gender}

Month = 0.4(Age) + 0.3(Gender) + U_{month} …

Decision = Month + 0.5(Age) + … + $U_{Decision}$



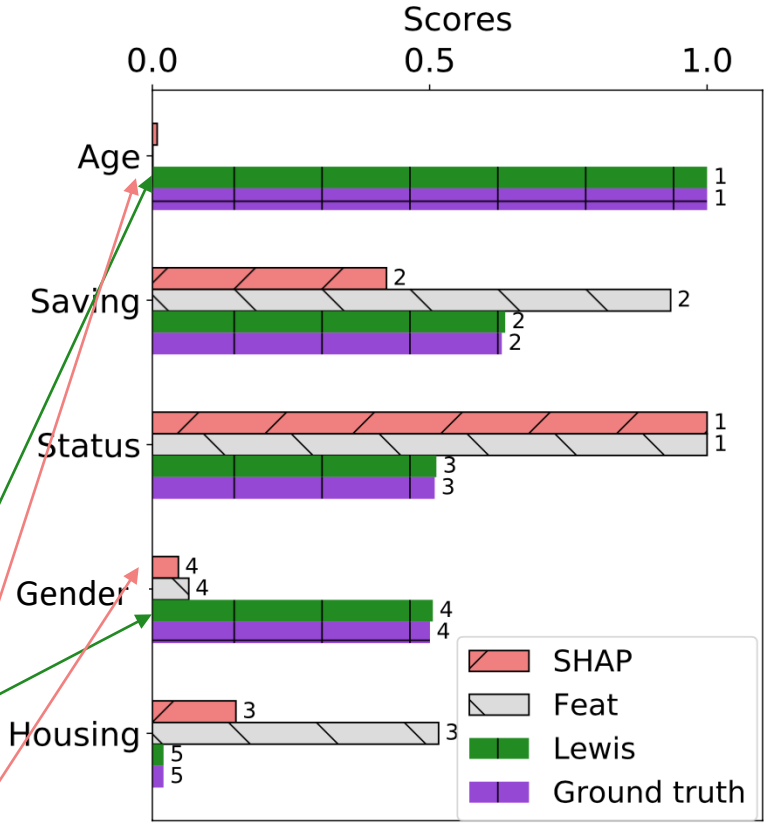
Experiments



Suppose the algorithm does not use gender and age for prediction

Age and gender still have a causal influence on the decision

SHAP picks up correlation between attributes– ranks age and gender the lowest



Synthetic dataset

Causal Shapley

- SHAP will be covered by student presentations – causal version
- Proposed a causal notion of the value function
- Captures direct and indirect effects of an attribute

Hekes et al, https://arxiv.org/abs/2011.01625
Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models

Common Challenges & Takeaways

- Dependence on causal structure
- Fragile with respect to noise in data
  - Distribution shift
- Need to adapt to evolving needs

- Causal explanations are translatable to real-world interventions
- Challenges
  - Require additional information
  - Computationally difficult
- Recourse
  - Helps individuals to reverse algorithm's decisions
  - Generally solved as an optimization problem

Data-based explanations

Explaining model outcomes

Rosa

Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data

ML Model

83%

Probability that Rosa's income < 50K

Why this output ?

Data-based explanations

Rosa

Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data

ML Model

83%

Probability that Rosa's income < 50K

Which part of the input data was most responsible for this prediction?

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

Idea: Select training data points to explain ML model behavior

Rosa

Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data

ML Model

83%

Probability that Rosa's income < 50K

These instances are the most responsible for Rosa's negative outcome

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

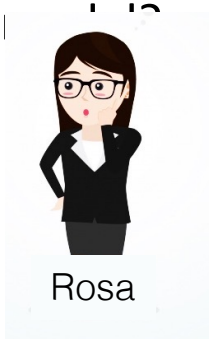
Approaches for Data-based Explanations

- **Objective:** debug data as a way to debug ML models, understand and explain model behavior and model predictions

- Popular approaches:
  - Data valuation
  - Influence functions
  - Representer points
  - Prototypes and criticisms

How valuable or influential is a data point?

- How much does each training data point contribute to the learned



Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data

Original Model

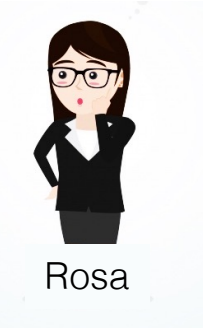
83%

Probability that Rosa's income < 50K

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

How valuable or influential is a data point?

- Leave-one-out approach



Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data

Updated Model

75%

Probability that Rosa's income < 50K

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
~~Nazia~~	~~36~~	~~Bachelors~~	~~Unmarried~~	…	~~Asian~~	~~female~~	~~40~~	~~<50K~~
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

Cook, R. D. Detection of influential observation in linear regression. Technometrics, 19(1):15–18, 1977

How valuable or influential is a data point?

- Leave-one-out approach

Original model performance

Updated model performance

Value (

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

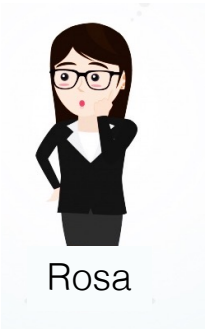
) = 0.83 – 0.75
= 0.08

- Might be expensive to compute for each data point
- Not reasonable when training data has duplicates

When is a data valuation approach good?

When is a data valuation approach good?

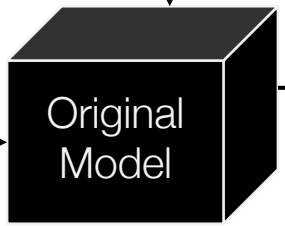
1. **Null Element:** If adding training data point t_1 to *any* subset of training data never changes the learned model's performance:



value(t_1) = 0

Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data



83%

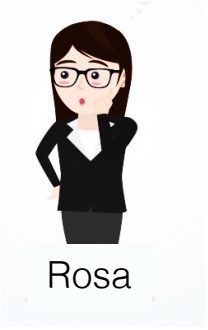
Probability that Rosa's income < 50K

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

Ghorbani, A., Zou, J.. Data Shapley: Equitable Valuation of Data for Machine Learning. ICML. 2019

When is a data valuation approach good?

1. **Null Element:** If adding training data point t_1 to *any* subset of training data never changes the learned model's performance:



Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data

Updated Model

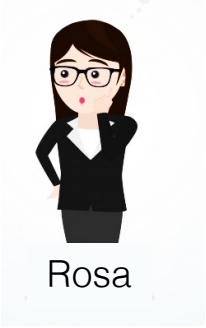
83%

Probability that Rosa's income < 50K

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
~~Matt~~	~~26~~	~~Bachelors~~	~~Married~~	~~…~~	~~White~~	~~male~~	~~40~~	~~≥50K~~
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

When is a data valuation approach good?

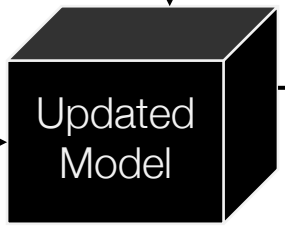
1. **Null Element:** If adding training data point t_1 to *any* subset of training data never changes the learned model's performance:



Rosa

Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data



83%

Probability that Rosa's income < 50K

Value (

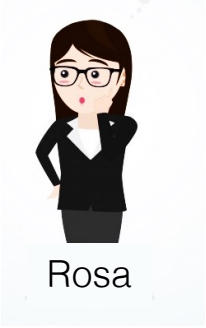
Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

) = 0

When is a data valuation approach good?

2. **Symmetry:** If adding training data point t_1 or t_2 to any subset of training data always results in the same change in performance:

$$value(t_1) = value(t_2)$$



Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data

Updated Model

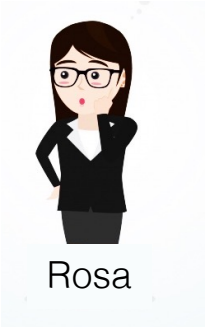
80%

Probability that Rosa's income < 50K

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
~~Yeji~~	~~50~~	~~Masters~~	~~Married~~	~~…~~	~~Asian~~	~~male~~	~~16~~	~~≥50K~~
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

When is a data valuation approach good?

2. Symmetry: If adding training data point t_1 or t_2 to any subset of training data always results in the same change in performance:



Rosa

Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data

Updated Model

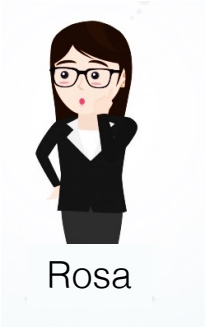
80%

Probability that Rosa's income < 50K

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
~~Neel~~	~~45~~	~~Masters~~	~~Unmarried~~	…	~~Black~~	~~male~~	~~28~~	~~<50K~~
…	…	…	…	…	…	…	…	…

When is a data valuation approach good?

2. Symmetry: If adding training data point t_1 or t_2 to any subset of training data always results in the same change in performance:



Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data

Updated Model

80%

Probability that Rosa's income < 50K

Value (| Name | Age | Education | Marital | … | Race | Gender | Hours | Income |)

= Value (| Name | Age | Education | Marital | … | Race | Gender | Hours | Income |)

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

When is a data valuation approach good?

3. **Linearity:** In ML, a performance metric (e.g., accuracy) is often the sum of the performance metric on individual test points

In this case, value of a datum should be sum of its value for each prediction

Value (

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

) = Value for 

+ Value for 

+ …

Other approaches: Data Shapley and Influence

- Given training data D and performance metric P, the value of a data point t_1 is given by Data Shapley Value

Marginal contribution of t_1

$$\text{Value}\ (t_1) = \sum_{\substack{\text{Subsets S}\subset D \\ \text{not containing } t_1}} \frac{P(S \cup t_1) - P(S)}{\binom{|D|-1}{|S|}}$$

All possible subsets

# of subsets of size |S| in D

Idea of *influence* from robust statistics
How do the model parameters change as we upweight z_{train} by an infinitesimal amount ϵ? Also group influence

Ghorbani, A., Zou, J.. Data Shapley: Equitable Valuation of Data for Machine Learning. ICML. 2019

Koh, P.W. & Liang, P. Understanding Black-box Predictions via Influence Functions. ICML 2017

Basu, S., You, X., and Feizi, S. On Second-Order Group Influence Functions for Black-Box Predictions. ICML 2020